Advanced Regression: Random effects and hierarchical models I

Garyfallos Konstantinoudis

Feb 20, 2024

Motivation

All methods presented so far assume that the observations are iid – **i**ndependent and **i**dentically **d**istributed

- **Independent**: The observations are conditionally independent from each other

$$cor(x_i, x_{i'}) = 0 \text{ for all } i, i' \in 1, ..., n$$

- **Identically**: All observations come from the same distribution. For example, from a Normal distribution with the same mean and variance.

Exchangeability: Allows for dependence between observations and only states that future observations behave like past ones.

Motivation: How realistic is iid?

- Often our data contains structure depending on how our data was sampled.
  - Within K boroughs in London we select n participants...
  - From K schools we sample n students...
  - From K hospitals we select n patients...

Grouping creates dependence: Observations within a group are likely to be more similar to each other than to observations from other groups.

Motivation: GP patient data

- We are interested in **the relationship between cholesterol and age**.
- We take measurements of patients from $K = 12$ GPs.

```
table(data_chol$doctor)
```

```
 1  2  3  4  5  6  7  8  9 10 11 12
36 36 36 39 36 36 39 36 36 39 36 36
```

```
head(data_chol)
```

```
# A tibble: 6 x 6
   chol doctor   age   bmi agedoc   sex
  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl>
1  7.13      1    54  27.4     55     0
2  7.7       1    55  29.1     55     0
3  7.3       1    56  27.9     55     0
4  6.89      1    71  26.7     55     1
5  6.9       1    72  26.7     55     1
6  7.9       1    73  29.7     55     1
```

Pooled analysis

$$y_i = \alpha_0 + \beta x_i + \epsilon_i$$

Assumptions: All observations independent (incorrect).

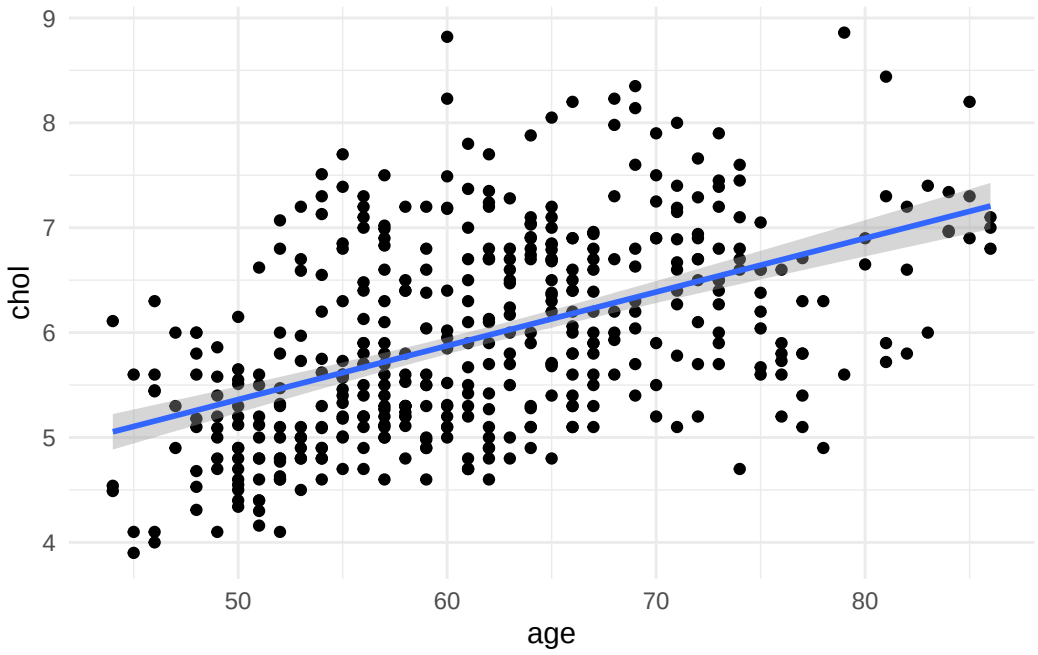
Consequences:

- Estimated errors on regression coefficients are too small.
- Overstate significance of association.

GP data: pooled analysis

```
model_pooled <- lm(chol ~ age, data = data_chol)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



GP data: pooled analysis

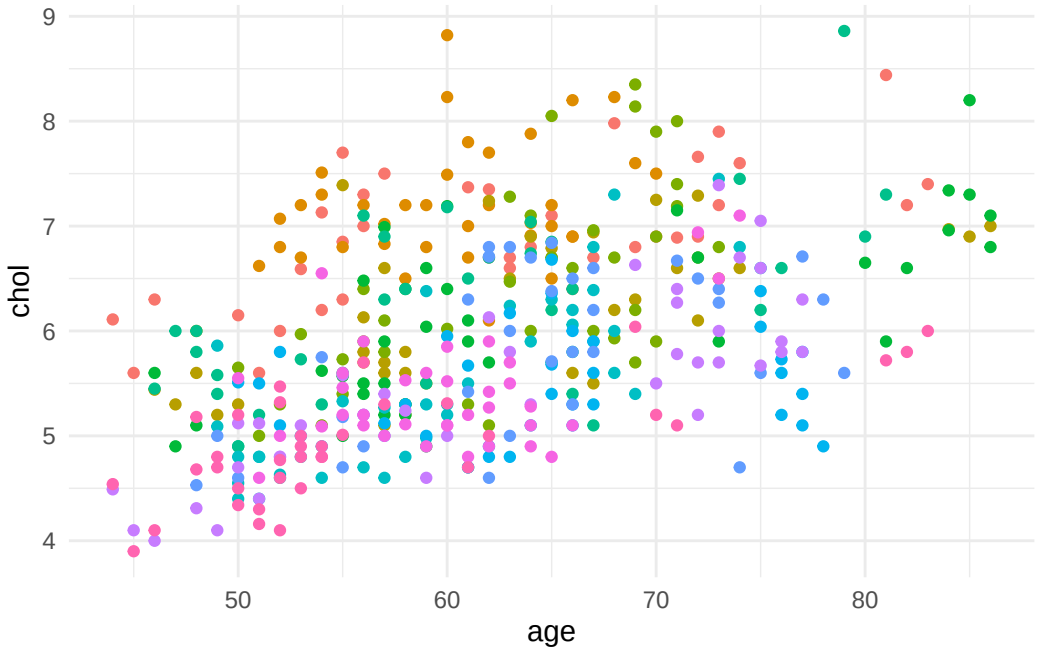
```
summary(model_pooled)
```

```
Call:
lm(formula = chol ~ age, data = data_chol)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8971 -0.6206 -0.1105  0.5693  2.9456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.798691   0.268571   10.42   <2e-16 ***
age         0.051262   0.004301   11.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8362 on 439 degrees of freedom
Multiple R-squared:  0.2445,    Adjusted R-squared:  0.2428
```

```
F-statistic: 142.1 on 1 and 439 DF,  p-value: < 2.2e-16
```

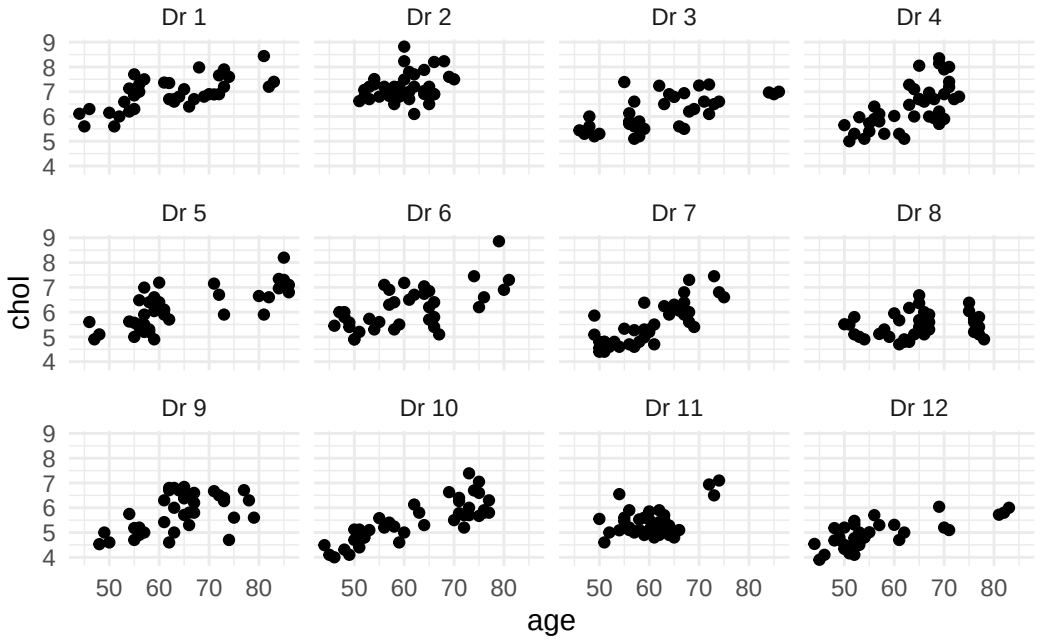
GP data: pooled analysis

```
data_chol |>
  mutate(doctor_name = factor(str_c("Dr ", doctor), levels = str_c("Dr ", 1:12))) |>
  ggplot(aes(x = age, y = chol, colour = doctor_name)) +
  geom_point() +
  theme_minimal() +
  theme(legend.position = "none")
```



GP data: pooled analysis

```
data_chol |>
  mutate(doctor_name = factor(str_c("Dr ", doctor), levels = str_c("Dr ", 1:12))) |>
  ggplot(aes(x = age, y = chol)) +
  geom_point() +
  facet_wrap(~doctor_name) +
  theme_minimal()
```



Accounting for dependence

When we ignore dependence:

- standard **errors too small**
- p−values too small / confidence intervals too narrow
- we **over-estimate significance**.

Intuitively, there is less information in the data than an independent sample.

Accounting for dependence

We can account for **dependence** by:

1. Perform analysis for each group separately.
2. Calculate summary measures for each group and use standard analysis (group-level analysis).
3. Fixed effects model to account for group structures.
4. Use random effects models that explicitly model the similarity of observations in a group.

Motivation: individual-level and group-level

Observations are **grouped** with grouping information known.

Multi-level: Multiple levels of groupings, e.g. classrooms within schools within districts.

Variables can be measured on the individual and group level.

1. Separate analysis

Estimate **separate regression coefficients** for each group.

Assumptions: Independence between groups.

Consequences:

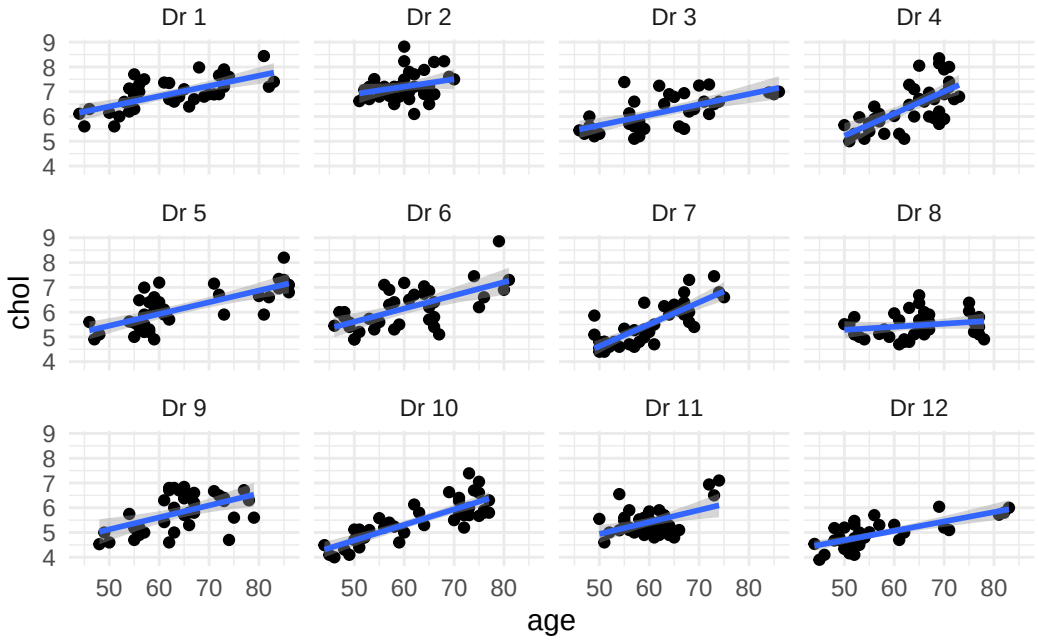
- This is a reasonable approach to exploratory analysis.
- If the number of individuals in each group is small, we will get imprecise estimates.
- Multiple testing is an issue.

1. Separate analysis

```
model_separate <- lm(chol ~ age | doctor, data = data_chol)
```

```
data_chol |>
  mutate(doctor_name = factor(str_c("Dr ", doctor), levels = str_c("Dr ", 1:12))) |>
  ggplot(aes(x = age, y = chol)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~doctor_name) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



2. Group-level analysis

Summarise outcome and predictors for each group k, e.g. using mean or median.

```
data_grouped <- data_chol |>
  group_by(doctor) |>
  summarise(chol_grouped = mean(chol), age_grouped = mean(age))
```

2. Group-level analysis

Treat the group summaries as observations

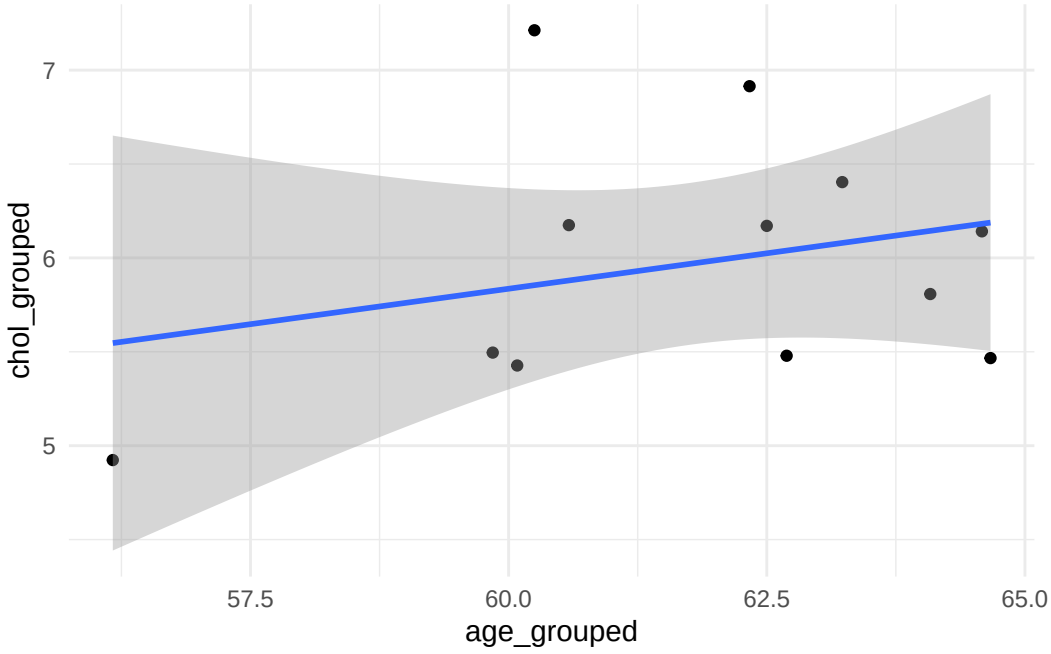
```
model_grouped <- lm(chol_grouped ~ age_grouped, data = data_grouped)
```

Consequences:

- One regression line fit: Associations between outcome and predictors are the same for each group.
- Independence between groups.
- All groups are treated equal, irrespective of size

2. Group-level analysis

```
`geom_smooth()` using formula = 'y ~ x'
```



2. Group-level analysis

```
summary(model_grouped)
```

```
Call:
lm(formula = chol_grouped ~ age_grouped, data = data_grouped)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7216 -0.4513 -0.1844  0.3020  1.3576

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.30687    5.05233   0.259    0.801
age_grouped  0.07548    0.08176   0.923    0.378
```

```
Residual standard error: 0.67 on 10 degrees of freedom
Multiple R-squared:  0.07854,   Adjusted R-squared:  -0.0136
F-statistic: 0.8524 on 1 and 10 DF,  p-value: 0.3776
```

2. Group-level analysis

- This model **lacks power** as the number of data points used is the number of groups ($k < n$)
- Regression coefficients will be **averaged over all groups** so real within-group effects may be diluted.
- Regression coefficients will only be significant if there are similar significant association effects across all groups.

Inverse variance weighted (IVW) meta-analysis

Each random variable is **weighted in inverse proportion to its variance**. Assume we have independent observations y_k with variance σ_k. Then the IVW estimate is defined as

$$\hat{y}_{\text{IVW}} = \frac{\sum_{k=1}^{K} y_k/\sigma_k}{\sum_{k=1}^{K} 1/\sigma_k}$$

Inverse variance weighted (IVW) meta-analysis

Weighted regression over groups

Assume y_k is a vector of group summaries, x_k is a $k \times p$ matrix of group summaries. Assume w is a diagonal matrix with $w[k,k] = \frac{1}{\sigma_k^2}$, then the **weighted least squares** estimate is defined as

$$\hat{\beta}_w = (x_k^t w x_k)^{-1} x_k^t w y_k$$

3. Fixed effects

Motivation:

- Keep the idea of modelling within groups, Allow **associations to differ across groups**.
- But now we model all the data (n observations) together: Maximise the power to detect associations.

3. Fixed effects

Joint model with group-specific intercept

$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

where α_k is a fixed effect.

- α_k captures the effect of unobserved group specific confounders.
- Residual errors ϵ_i are assumed independent.

3. Fixed effects

A fixed effects model is fit in the same way as the simple linear model including the group as a covariate.

Assumptions: Information on k comes from observations in group k only.

Consequences:

- By including group effects we adjust for group characteristics.
- But introduces a number of parameters (one for each group).
- May be a problem if there are few observations in some groups.

3. Fixed effects with `lm()`

There are two different types of fixed effect:

1. Group-specific intercept α_k
$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

2. Group-specific slope β_k
$$y_i = \alpha_0 + \beta_k x_i + \epsilon_i$$

Varying intercept with `lm()`

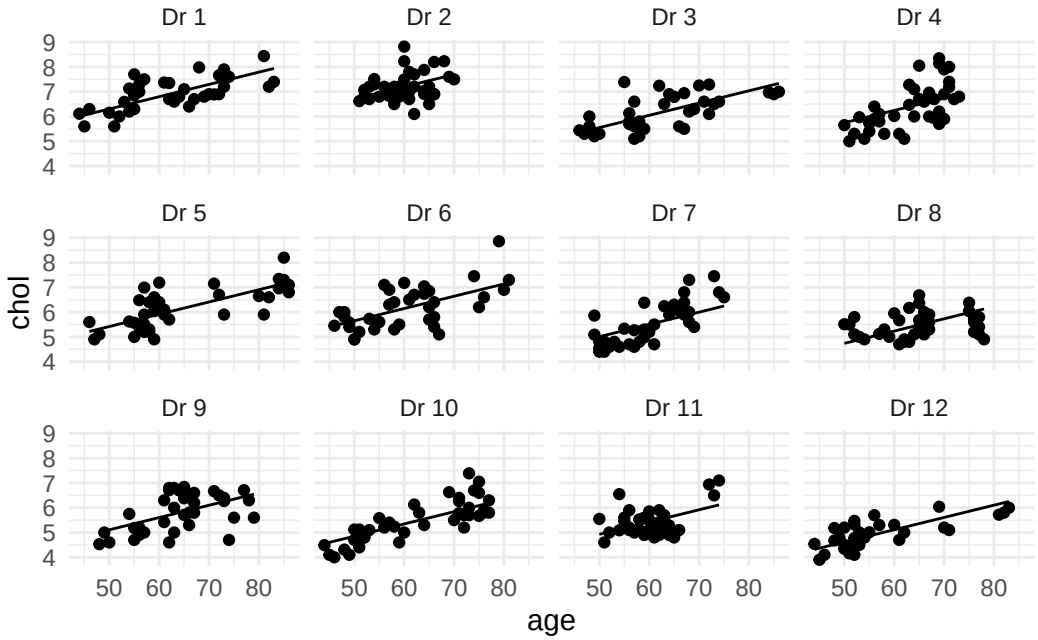
1. Group-specific intercept α_k
$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

```r
model_varying_intercept <- lm(chol ~ as.factor(doctor) + age, data = data_chol)
```

Varying intercept with `lm()`

```r
summary(model_varying_intercept)
```

```
Call:
lm(formula = chol ~ as.factor(doctor) + age, data = data_chol)

Residuals:
     Min       1Q   Median       3Q      Max
-1.59881 -0.40321 -0.08463  0.37929  1.77313

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.826236   0.213854  17.892  < 2e-16 ***
as.factor(doctor)2   0.400993   0.136014   2.948  0.00337 **
as.factor(doctor)3  -0.752146   0.135865  -5.536 5.41e-08 ***
as.factor(doctor)4  -0.555317   0.133254  -4.167 3.73e-05 ***
as.factor(doctor)5  -0.884528   0.136039  -6.502 2.21e-10 ***
as.factor(doctor)6  -0.653299   0.135970  -4.805 2.15e-06 ***
as.factor(doctor)7  -1.295580   0.133444  -9.709  < 2e-16 ***
as.factor(doctor)8  -1.563657   0.136053 -11.493  < 2e-16 ***
as.factor(doctor)9  -1.193645   0.135970  -8.779  < 2e-16 ***
as.factor(doctor)10 -1.453255   0.133231 -10.908  < 2e-16 ***
as.factor(doctor)11 -1.376027   0.136039 -10.115  < 2e-16 ***
as.factor(doctor)12 -1.685593   0.137173 -12.288  < 2e-16 ***
age                  0.049543   0.003065  16.164  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5764 on 428 degrees of freedom
Multiple R-squared:   0.65, Adjusted R-squared:  0.6402
F-statistic: 66.24 on 12 and 428 DF,  p-value: < 2.2e-16
```



Varying intercept with `lm()`

Varying slope with `lm()`

2. Group-specific slope β_k

$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

```
model_varying_slope <- lm(chol ~ age:as.factor(doctor), data = data_chol)
```

Varying slope with `lm()`

```
summary(model_varying_slope)
```

```
Call:
lm(formula = chol ~ age:as.factor(doctor), data = data_chol)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-1.56821 -0.41837 -0.07627  0.38652  1.67691

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.845418   0.193641  14.694   <2e-16 ***
age:as.factor(doctor)1   0.064655   0.003392  19.059   <2e-16 ***
age:as.factor(doctor)2   0.072176   0.003571  20.210   <2e-16 ***
age:as.factor(doctor)3   0.052904   0.003382  15.644   <2e-16 ***
age:as.factor(doctor)4   0.056631   0.003365  16.831   <2e-16 ***
age:as.factor(doctor)5   0.050906   0.003247  15.676   <2e-16 ***
age:as.factor(doctor)6   0.054907   0.003492  15.722   <2e-16 ***
age:as.factor(doctor)7   0.044971   0.003540  12.703   <2e-16 ***
age:as.factor(doctor)8   0.040084   0.003301  12.143   <2e-16 ***
age:as.factor(doctor)9   0.046254   0.003333  13.877   <2e-16 ***
age:as.factor(doctor)10 0.042601   0.003337  12.765   <2e-16 ***
age:as.factor(doctor)11 0.043010   0.003577  12.025   <2e-16 ***
age:as.factor(doctor)12 0.037020   0.003750   9.873   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5817 on 428 degrees of freedom
Multiple R-squared:  0.6435,    Adjusted R-squared:  0.6336
F-statistic: 64.39 on 12 and 428 DF,  p-value: < 2.2e-16
```

Varying slope with `lm()`

Fixed effects with `lm()`

Main formula: `y  x`, where `y` is the outcome and `x` the predictor(s)

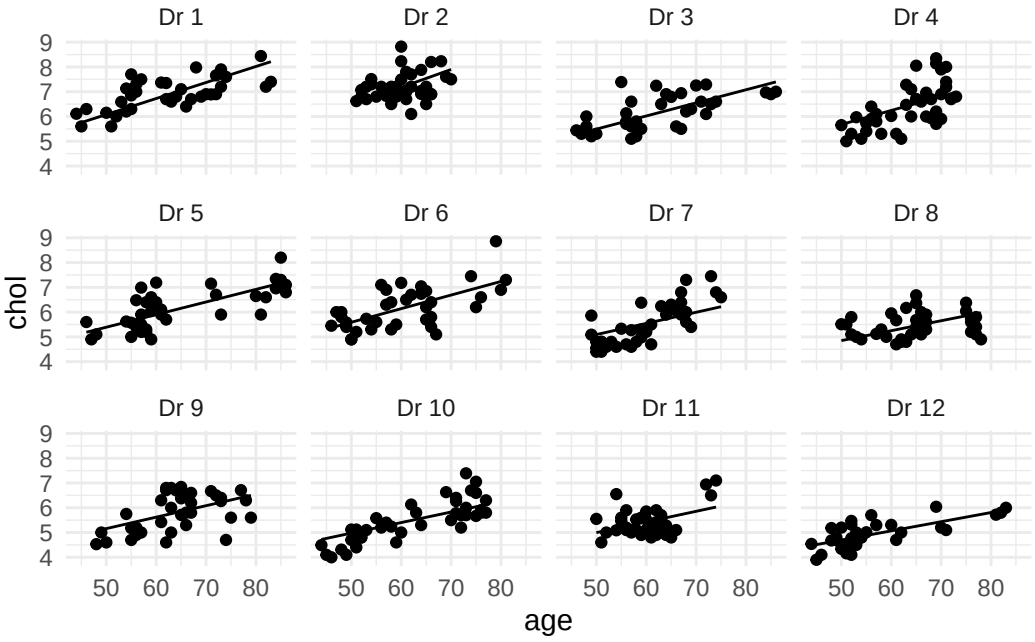
Predictors can be added as

`+` | main effect |
`:` | interaction only |
`*` | main effect and interaction |

Use `summary()`, `coef()` and `fitted()` to get values.

Fixed effects: Disadvantages

- Fixed effects account for **any** unobserved group-specific confounders, so including both a group-specific intercept and slope is not identifiable.



  - When the intercept α_k is group-specific, then the slope is assumed to be the same for all groups.
  - When slope β_k is group-specific, then the intercept is assumed to be the same for all groups.

Fixed effects: Disadvantages

- If we add new groups to the dataset we may not consistently estimate α_k:
  - Consider α_1, the intercept for the first group.
  - When we add new groups, the slope may vary.
  - Changing slope will change the intercept, also α_1.
- Information on α_k or β_k comes only from observations in group k and we need to estimate one parameter per group.

Take away: Structured Data

- Most statistical methods are developed for independent and identically distributed (iid) data, but often in practice we observe structured data, where **there is an intrinsic group structure**.

- Grouping creates dependence: Observations within a group are likely to be more similar to each other than to observations from other groups.
- Ignoring the group structure can lead to over-confident results or even false positives.
- Analysing each group separately, we do not assume any shared mechanisms and need to fit a model on the samples within a group only. Aggregating and working only on the group-level drastically reduces the sample size k.